A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation

Stephen Tratz and **Eduard Hovy**
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
{stratz,hovy}@isi.edu

Abstract

The automatic interpretation of noun-noun compounds is an important subproblem within many natural language processing applications and is an area of increasing interest. The problem is difficult, with disagreement regarding the number and nature of the relations, low inter-annotator agreement, and limited annotated data. In this paper, we present a novel taxonomy of relations that integrates previous relations, the largest publicly-available annotated dataset, and a supervised classification method for automatic noun compound interpretation.

1 Introduction

Noun compounds (e.g., 'maple leaf') occur very frequently in text, and their interpretation—determining the relationships between adjacent nouns as well as the hierarchical dependency structure of the NP in which they occur—is an important problem within a wide variety of natural language processing (NLP) applications, including machine translation (Baldwin and Tanaka, 2004) and question answering (Ahn et al., 2005). The interpretation of noun compounds is a difficult problem for various reasons (Spärck Jones, 1983). Among them is the fact that no set of relations proposed to date has been accepted as complete and appropriate for general-purpose text. Regardless, automatic noun compound interpretation is the focus of an upcoming SEMEVAL task (Butnariu et al., 2009).

Leaving aside the problem of determining the dependency structure among strings of three or more nouns—a problem we do not address in this paper—automatic noun compound interpretation requires a taxonomy of noun-noun relations, an automatic method for accurately assigning the relations to noun compounds, and, in the case of supervised classification, a sufficiently large dataset for training.

Earlier work has often suffered from using taxonomies with coarse-grained, highly ambiguous predicates, such as prepositions, as various labels (Lauer, 1995) and/or unimpressive inter-annotator agreement among human judges (Kim and Baldwin, 2005). In addition, the datasets annotated according to these various schemes have often been too small to provide wide coverage of the noun compounds likely to occur in general text.

In this paper, we present a large, fine-grained taxonomy of 43 noun compound relations, a dataset annotated according to this taxonomy, and a supervised, automatic classification method for determining the relation between the head and modifier words in a noun compound. We compare and map our relations to those in other taxonomies and report the promising results of an inter-annotator agreement study as well as an automatic classification experiment. We examine the various features used for classification and identify one very useful, novel family of features. Our dataset is, to the best of our knowledge, the largest noun compound dataset yet produced. We will make it available via http://www.isi.edu.

2 Related Work

2.1 Taxonomies

The relations between the component nouns in noun compounds have been the subject of various linguistic studies performed throughout the years, including early work by Jespersen (1949). The taxonomies they created are varied. Lees created an early taxonomy based primarily upon grammar (Lees, 1960). Levi's influential work postulated that *complex nominals* (Levi's name for noun compounds that also permits certain adjectival modifiers) are all derived either via nominalization or

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 678–687,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

by deleting one of nine predicates (i.e., CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM, ABOUT) from an underlying sentence construction (Levi, 1978). Of the taxonomies presented by purely linguistic studies, our categories are most similar to those proposed by Warren (1978), whose categories (e.g., MATERIAL+ARTEFACT, OBJ+PART) are generally less ambiguous than Levi's.

In contrast to studies that claim the existence of a relatively small number of semantic relations, Downing (1977) presents a strong case for the existence of an unbounded number of relations. While we agree with Downing's belief that the number of relations is unbounded, we contend that the vast majority of noun compounds fits within a relatively small set of categories.

The relations used in computational linguistics vary much along the same lines as those proposed earlier by linguists. Several lines of work (Finin, 1980; Butnariu and Veale, 2008; Nakov, 2008) assume the existence of an unbounded number of relations. Others use categories similar to Levi's, such as Lauer's (1995) set of prepositional paraphrases (i.e., OF, FOR, IN, ON, AT, FROM, WITH, ABOUT) to analyze noun compounds. Some work (e.g., Barker and Szpakowicz, 1998; Nastase and Szpakowicz, 2003; Girju et al., 2005; Kim and Baldwin, 2005) use sets of categories that are somewhat more similar to those proposed by Warren (1978). While most of the noun compound research to date is not domain specific, Rosario and Hearst (2001) create and experiment with a taxonomy tailored to biomedical text.

2.2 Classification

The approaches used for automatic classification are also varied. Vanderwende (1994) presents one of the first systems for automatic classification, which extracted information from online sources and used a series of rules to rank a set of most likely interpretations. Lauer (1995) uses corpus statistics to select a prepositional paraphrase. Several lines of work, including that of Barker and Szpakowicz (1998), use memory-based methods. Kim and Baldwin (2005) and Turney (2006) use nearest neighbor approaches based upon WordNet (Fellbaum, 1998) and Turney's Latent Relational Analysis, respectively. Rosario and Hearst (2001) utilize neural networks to classify compounds according to their domain-specific relation taxonomy. Moldovan et al. (2004) use SVMs as well as a novel algorithm (i.e., semantic scattering). Nastase et al. (2006) experiment with a variety of classification methods including memory-based methods, SVMs, and decision trees. Ó Séaghdha and Copestake (2009) use SVMs and experiment with kernel methods on a dataset labeled using a relatively small taxonomy. Girju (2009) uses cross-linguistic information from parallel corpora to aid classification.

3 Taxonomy

3.1 Creation

Given the heterogeneity of past work, we decided to start fresh and build a new taxonomy of relations using naturally occurring noun pairs, and then compare the result to earlier relation sets. We collected 17509 noun pairs and over a period of 10 months assigned one or more relations to each, gradually building and refining our taxonomy. More details regarding the dataset are provided in Section 4.

The relations we produced were then compared to those present in other taxonomies (e.g., Levi, 1978; Warren, 1978; Barker and Szpakowicz, 1998; Girju et al., 2005), and they were found to be fairly similar. We present a detailed comparison in Section 3.4.

We tested the relation set with an initial inter-annotator agreement study (our latest inter-annotator agreement study results are presented in Section 6). However, the mediocre results indicated that the categories and/or their definitions needed refinement. We then embarked on a series of changes, testing each generation by annotation using Amazon's Mechanical Turk service, a relatively quick and inexpensive online platform where requesters may publish tasks for anonymous online workers (Turkers) to perform. Mechanical Turk has been previously used in a variety of NLP research, including recent work on noun compounds by Nakov (2008) to collect short phrases for linking the nouns within noun compounds.

For the Mechanical Turk annotation tests, we created five sets of 100 noun compounds from noun compounds automatically extracted from a random subset of New York Times articles written between 1987 and 2007 (Sandhaus, 2008). Each of these sets was used in a separate annotation round. For each round, a set of 100 noun compounds was uploaded along with category defini-

Category Name	%	Example	Approximate Mappings
Causal Group			
Communicator of Communication	0.77	court order	$\supset$BGN:Agent, $\supset$L:Act$_a$+Product$_a$, $\supset$V:Subj
Performer of Act/Activity	2.07	police abuse	$\supset$BGN:Agent, $\supset$L:Act$_a$+Product$_a$, $\supset$V:Subj
Creator/Provider/Cause Of	2.55	ad revenue	$\subset$BGV:Cause(d-by), $\subset$L:Cause$_2$, $\subset$N:Effect
Purpose/Activity Group			
Perform/Engage_In	13.24	cooking pot	$\supset$BGV:Purpose, $\supset$L:For, $\approx$N:Purpose, $\supset$W:Activity$\cup$Purpose
Create/Provide/Sell	8.94	nicotine patch	∞BV:Purpose, $\subset$BG:Result, ∞G:Make-Produce, $\subset$GNV:Cause(s), ∞L:Cause$_1$$\cup$Make$_1$$\cup$For, $\subset$N:Product, $\supset$W:Activity$\cup$Purpose
Obtain/Access/Seek	1.50	shrimp boat	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Modify/Process/Change	1.50	eye surgery	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Mitigate/Oppose/Destroy	2.34	flak jacket	$\supset$BGV:Purpose, $\supset$L:For, $\approx$N:Detraction, $\supset$W:Activity$\cup$Purpose
Organize/Supervise/Authority	4.82	ethics board	$\supset$BGNV:Purpose/Topic, $\supset$L:For/About$_a$, $\supset$W:Activity
Propel	0.16	water gun	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Protect/Conserve	0.25	screen saver	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Transport/Transfer/Trade	1.92	freight train	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Traverse/Visit	0.11	tree traversal	$\supset$BGNV:Purpose, $\supset$L:For, $\supset$W:Activity$\cup$Purpose
Ownership, Experience, Employment, and Use			
Possessor + Owned/Possessed	2.11	family estate	$\supset$BGNVW:Possess*, $\supset$L:Have$_2$
Experiencer + Coginition/Mental	0.45	voter concern	$\supset$BNVW:Possess*, $\approx$G:Experiencer, $\supset$L:Have$_2$
Employer + Employee/Volunteer	2.72	team doctor	$\supset$BGNVW:Possess*, $\supset$L:For/Have$_2$, $\supset$BGN:Beneficiary
Consumer + Consumed	0.09	cat food	$\supset$BGNVW:Purpose, $\supset$L:For, $\supset$BGN:Beneficiary
User/Recipient + Used/Received	1.02	voter guide	$\supset$BNVW:Purpose, $\supset$G:Recipient, $\supset$L:For, $\supset$BGN:Beneficiary
Owned/Possessed + Possession	1.20	store owner	$\approx$G:Possession, $\supset$L:Have$_1$, $\approx$W:Belonging-Possessor
Experience + Experiencer	0.27	fire victim	$\approx$G:Experiencer, ∞L:Have$_1$
Thing Consumed + Consumer	0.41	fruit fly	$\supset$W:Obj-SingleBeing
Thing/Means Used + User	1.96	faith healer	$\approx$BNV:Instrument, $\approx$G:Means$\cup$Instrument, $\approx$L:Use, $\subset$W:MotivePower-Obj
Temporal Group			
Time [Span] + X	2.35	night work	$\approx$BNV:Time(At), $\supset$G:Temporal, $\approx$L:In$_c$, $\approx$W:Time-Obj
X + Time [Span]	0.50	birth date	$\supset$G:Temporal, $\approx$W:Obj-Time
Location and Whole+Part/Member of			
Location/Geographic Scope of X	4.99	hillside home	$\approx$BGV:Locat(ion/ive), $\approx$L:In$_a$$\cup$From$_b$, B:Source, $\approx$N:Location(At/From), $\approx$W:Place-Obj$\cup$PlaceOfOrigin
whole + part/member of	1.75	robot arm	$\supset$B:Possess*, $\approx$G:Part-Whole, $\supset$L:Have$_2$, $\approx$N:Part, $\approx$V:Whole-Part, $\approx$W:Obj-Part$\cup$Group-Member
Composition and Containment Group			
Substance/Material/Ingredient + Whole	2.42	plastic bag	$\subset$BNVW:Material*, ∞GN:Source, ∞L:From$_a$, $\approx$L:Have$_1$, ∞L:Make$_{2b}$, ∞N:Content
Part/Member + Collection/Config/Series	1.78	truck convoy	$\approx$L:Make$_{2ac}$, $\approx$N:Whole, $\approx$V:Part-Whole, $\approx$W:Parts-Whole
X + Spatial Container/Location/Bounds	1.39	shoe box	$\supset$B:Content$\cup$Located, $\supset$L:For, $\supset$L:Have$_1$, $\approx$N:Location, $\approx$W:Obj-Place
Topic Group			
Topic of Communication/Imagery/Info	8.37	travel story	$\supset$BGNV:Topic, $\supset$L:About$_{ab}$, $\supset$W:SubjectMatter, $\subset$G:Depiction
Topic of Plan/Deal/Arrangement/Rules	4.11	loan terms	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
Topic of Observation/Study/Evaluation	1.71	job survey	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
Topic of Cognition/Emotion	0.58	jazz fan	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
Topic of Expert	0.57	policy wonk	$\supset$BGNV:Topic, $\supset$L:About$_a$, $\supset$W:SubjectMatter
Topic of Situation	1.64	oil glut	$\supset$BGNV:Topic, $\approx$L:About$_c$
Topic of Event/Process	1.09	lava flow	$\supset$G:Theme, $\supset$V:Subj
Attribute Group			
Topic/Thing + Attrib	4.13	street name	$\supset$BNV:Possess*, $\approx$G:Property, $\supset$L:Have$_2$, $\approx$W:Obj-Quality
Topic/Thing + Attrib Value Charac Of	0.31	earth tone	
Attributive and Coreferential			
Coreferential	4.51	fighter plane	$\approx$BV:Equative, $\supset$G:Type$\cup$IS-A, $\approx$L:BE$_{bcd}$, $\approx$N:Type$\cup$Equality, $\approx$W:Copula
Partial Attribute Transfer	0.69	skeleton crew	$\approx$W:Resemblance, $\supset$G:Type
Measure + Whole	4.37	hour meeting	$\approx$G:Measure, $\subset$N:TimeThrough$\cup$Measure, $\approx$W:Size-Whole
Other			
Highly Lexicalized / Fixed Pair	0.65	pig iron	
Other	1.67	contact lens	

Table 1: The semantic relations, their frequency in the dataset, examples, and *approximate* relation mappings to previous relation sets. $\approx$-approximately equivalent; $\supset$/$\subset$-super/sub set; ∞-some overlap; $\cup$-union; initials BGLNVW refer respectively to the works of (Barker and Szpakowicz, 1998; Girju et al., 2005; Girju, 2007; Levi, 1978; Nastase and Szpakowicz, 2003; Vanderwende, 1994; Warren, 1978).

tions and examples. Turkers were asked to select one or, if they deemed it appropriate, two categories for each noun pair. After all annotations for the round were completed, they were examined, and any taxonomic changes deemed appropriate (e.g., the creation, deletion, and/or modification of categories) were incorporated into the taxonomy before the next set of 100 was uploaded. The categories were substantially modified during this process. They are shown in Table 1 along with examples and an approximate mapping to several other taxonomies.

3.2 Category Descriptions

Our categories are defined with sentences. For example, the SUBSTANCE category has the definition n_1 *is one of the primary physical substances/materials/ingredients that* n_2 *is made/composed out of/from.* Our LOCATION category's definition reads n_1 *is the location / geographic scope where* n_2 *is at, near, from, generally found, or occurs.* Defining the categories with sentences is advantageous because it is possible to create straightforward, explicit defintions that humans can easily test examples against.

3.3 Taxonomy Groupings

In addition to influencing the category definitions, some taxonomy groupings were altered with the hope that this would improve inter-annotator agreement for cases where Turker disagreement was systematic. For example, LOCATION and WHOLE + PART/MEMBER OF were commonly disagreed upon by Turkers so they were placed within their own taxonomic subgroup. The ambiguity between these categories has previously been observed by Girju (2009).

Turkers also tended to disagree between the categories related to composition and containment. Due this apparent similarity they were also grouped together in the taxonomy.

The ATTRIBUTE categories are positioned near the TOPIC group because some Turkers chose a TOPIC category when an ATTRIBUTE category was deemed more appropriate. This may be because attributes are relatively abstract concepts that are often somewhat *descriptive* of whatever possesses them. A prime example of this is *street name*.

3.4 Contrast with other Taxonomies

In order to ensure completeness, we mapped into our taxonomy the relations proposed in most previous work including those of Barker and Szpakowicz (1998) and Girju et al. (2005). The results, shown in Table 1, demonstrate that our taxonomy is similar to several taxonomies used in other work. However, there are three main differences and several less important ones. The first major difference is the absence of a significant THEME or OBJECT category. The second main difference is that our taxonomy does not include a PURPOSE category and, instead, has several smaller categories. Finally, instead of possessing a single TOPIC category, our taxonomy has several, finer-grained TOPIC categories. These differences are significant because THEME/OBJECT, PURPOSE, and TOPIC are typically among the most frequent categories.

THEME/OBJECT is typically the category to which other researchers assign noun compounds whose head noun is a nominalized verb and whose modifier noun is the THEME/OBJECT of the verb. This is typically done with the justification that the relation/predicate (the root verb of the nominalization) is overtly expressed.

While including a THEME/OBJECT category has the advantage of simplicity, its disadvantages are significant. This category leads to a significant ambiguity in examples because many compounds fitting the THEME/OBJECT category also match some other category as well. Warren (1978) gives the examples of *soup pot* and *soup container* to illustrate this issue, and Girju (2009) notes a substantial overlap between THEME and MAKE-PRODUCE. Our results from Mechanical Turk showed significant overlap between PURPOSE and OBJECT categories (present in an earlier version of the taxonomy). For this reason, we do not include a separate THEME/OBJECT category. If it is important to know whether the modifier also holds a THEME/OBJECT relationship, we suggest treating this as a separate classification task.

The absence of a single PURPOSE category is another distinguishing characteristic of our taxonomy. Instead, the taxonomy includes a number of finer-grained categories (e.g., PERFORM/ENGAGE_IN), which can be conflated to create a PURPOSE category if necessary. During our Mechanical Turk-based refinement process, our now-defunct PURPOSE category was found to be ambiguous with many other categories as well as difficult to define. This problem has been noted by others. For example, Warren (1978) points out that *tea* in *tea cup* qualifies as both the *content* and the *purpose* of the *cup*. Similarly, while WHOLE+PART/MEMBER was selected by most Turkers for *bike tire*, one individual chose PURPOSE. Our investigation identified five main purpose-like relations that most of our PURPOSE examples can be divided into, including activity performance (PERFORM/ENGAGE_IN), creation/provision (CREATE/PROVIDE/CAUSE OF), obtainment/access (OBTAIN/ACCESS/SEEK), supervision/management (ORGANIZE/SUPERVISE/AUTHORITY), and opposition (MITIGATE/OPPOSE/DESTROY).

The third major distinguishing different between our taxonomy and others is the absence of a single TOPIC/ABOUT relation. Instead, our taxonomy has several finer-grained categories that can be conflated into a TOPIC category. Unlike the previous two distinguishing characteristics, which were motivated primarily by Turker annotations, this separation was largely motivated by author dissatisfaction with a single TOPIC category.

Two differentiating characteristics of less importance are the absence of BENEFICIARY or SOURCE categories (Barker and Szpakowicz, 1998; Nastase and Szpakowicz, 2003; Girju et al., 2005). Our EMPLOYER, CONSUMER, and USER/RECIPIENT categories combined more or less cover BENEFICIARY. Since SOURCE is ambiguous in multiple ways including causation (*tsunami injury*), provision (*government grant*), ingredients (*rice wine*), and locations (*north wind*), we chose to exclude it.

4 Dataset

Our noun compound dataset was created from two principal sources: an in-house collection of terms extracted from a large corpus using part-of-speech tagging and mutual information and the Wall Street Journal section of the Penn Treebank. Compounds including one or more proper nouns were ignored. In total, the dataset contains 17509 unique, out-of-context examples, making it by far the largest hand-annotated compound noun dataset in existence that we are aware of. Proper nouns were not included.

The next largest available datasets have a variety of drawbacks for noun compound interpretation in general text. Kim and Baldwin's (2005) dataset is the second largest available dataset, but inter-annotator agreement was only 52.3%, and the annotations had an usually lopsided distribution; 42% of the data has TOPIC labels. Most (73.23%) of Girju's (2007) dataset consists of noun-preposition-noun constructions. Rosario and Heart's (2001) dataset is specific to the biomedical domain, while Ó Séaghdha and Copestake's (2009) data is labeled with only 5 extremely coarse-grained categories. The remaining datasets are too small to provide wide coverage. See Table 2 below for size comparison with other publicly available, semantically annotated datasets.

Size	Work
17509	Tratz and Hovy, 2010
2169	Kim and Baldwin, 2005
2031	*Girju, 2007*
1660	Rosario and Hearst, 2001
1443	Ó Séaghdha and Copestake, 2007
505	*Barker and Szpakowicz, 1998*
600	*Nastase and Szpakowicz, 2003*
395	Vanderwende, 1994
385	Lauer, 1995

Table 2: Size of various available noun compound datasets labeled with relation annotations. Italics indicate that the dataset contains n-prep-n constructions and/or non-nouns.

5 Automated Classification

We use a Maximum Entropy (Berger et al., 1996) classifier with a large number of boolean features, some of which are novel (e.g., the inclusion of words from WordNet definitions). Maximum Entropy classifiers have been effective on a variety of NLP problems including preposition sense disambiguation (Ye and Baldwin, 2007), which is somewhat similar to noun compound interpretation. We use the implementation provided in the MALLET machine learning toolkit (McCallum, 2002).

5.1 Features Used

WordNet-based Features

- {Synonyms, Hypernyms} for all NN and VB entries for each word
- Intersection of the words' hypernyms
- All terms from the 'gloss' for each word
- Intersection of the words' 'gloss' terms
- Lexicographer file names for each word's NN and VB entries (e.g., n_1:substance)

- Logical AND of lexicographer file names for the two words (e.g., n_1:substance $\wedge$ n_2:artifact)
- Lists of all link types (e.g., meronym links) associated with each word
- Logical AND of the link types (e.g., n_1:hasMeronym(s) $\wedge$ n_2:hasHolonym(s))
- Part-of-speech (POS) indicators for the existence of VB, ADJ, and ADV entries for each of the nouns
- Logical AND of the POS indicators for the two words
- 'Lexicalized' indicator for the existence of an entry for the compound as a single term
- Indicators if either word is a part of the other word according to Part-Of links
- Indicators if either word is a hypernym of the other
- Indicators if either word is in the definition of the other

Roget's Thesaurus-based Features

- Roget's divisions for all noun (and verb) entries for each word
- Roget's divisions shared by the two words

Surface-level Features

- Indicators for the suffix types (e.g., de-adjectival, de-nominal [non]agentive, de-verbal [non]agentive)
- Indicators for degree, number, order, or locative prefixes (e.g., ultra-, poly-, post-, and inter-, respectively)
- Indicators for whether or not a preposition occurs within either term (e.g., 'down' in 'breakdown')
- The last {two, three} letters of each word

Web 1T N-gram Features

To provide information related to term usage to the classifier, we extracted trigram and 4-gram features from the Web 1T Corpus (Brants and Franz, 2006), a large collection of n-grams and their counts created from approximately one trillion words of Web text. Only n-grams containing lowercase words were used. 5-grams were not used due to memory limitations. Only n-grams containing both terms (including plural forms) were extracted. Table 3 describes the extracted n-gram features.

5.2 Cross Validation Experiments

We performed 10-fold cross validation on our dataset, and, for the purpose of comparison, we also performed 5-fold cross validation on Ó Séaghdha's (2007) dataset using his folds. Our classification accuracy results are 79.3% on our data and 63.6% on the Ó Séaghdha data. We used the χ^2 measure to limit our experiments to the most useful 35000 features, which is the point where we obtain the highest results on Ó Séaghdha's data. The 63.6% figure is similar to the best previously reported accuracy for this dataset of 63.1%, which was obtained by Ó Séaghdha and Copestake (2009) using kernel methods.

For comparison with SVMs, we used Thorsten Joachims' $SVM^{multiclass}$, which implements an optimization solution to Cramer and Singer's (2001) multiclass SVM formulation. The best results were similar, with 79.4% on our dataset and 63.1% on Ó Séaghdha's. $SVM^{multiclass}$ was, however, observed to be very sensitive to the tuning of the C parameter, which determines the tradeoff between training error and margin width. The best results for the datasets were produced with C set to 5000 and 375 respectively.

Trigram Feature Extraction Patterns			
text	<n_1>	<n_2>	
<*>	<n_1>	<n_2>	
<n_1>	<n_2>	text	
<n_1>	<n_2>	<*>	
<n_1>	text	<n_2>	
<n_2>	text	<n_1>	
<n_1>	<*>	<n_2>	
<n_2>	<*>	<n_1>	
4-Gram Feature Extraction Patterns			
<n_1>	<n_2>	text	text
<n_1>	<n_2>	<*>	text
text	<n_1>	<n_2>	text
text	text	<n_1>	<n_2>
text	<*>	<n_1>	<n_2>
<n_1>	text	text	<n_2>
<n_1>	text	<*>	<n_2>
<n_1>	<*>	text	<n_2>
<n_1>	<*>	<*>	<n_2>
<n_2>	text	text	<n_1>
<n_2>	text	<*>	<n_1>
<n_2>	<*>	text	<n_1>
<n_2>	<*>	<*>	<n_1>

Table 3: Patterns for extracting trigram and 4-Gram features from the Web 1T Corpus for a given noun compound (n_1 n_2).

To assess the impact of the various features, we ran the cross validation experiments for each feature type, alternating between including *only* one feature type and including all feature types *except* that one. The results for these runs using the Maximum Entropy classifier are presented in Table 4.

There are several points of interest in these results. The WordNet gloss terms had a surprisingly strong influence. In fact, by themselves they proved roughly as useful as the hypernym features, and their removal had the single strongest negative impact on accuracy for our dataset. As far as we know, this is the first time that WordNet definition words have been used as features for noun compound interpretation. In the future, it may be valuable to add definition words from other machine-readable dictionaries. The influence of the Web 1T n-gram features was somewhat mixed. They had a positive impact on the Ó Séaghdha data, but their affect upon our dataset was limited and mixed, with the removal of the 4-gram features actually improving performance slightly.

	Our Data		Ó Séaghdha Data	
	1	M-1	1	M-1
WordNet-based				
synonyms	0.674	0.793	0.469	0.626
hypernyms	0.753	0.787	0.539	0.626
hypernyms$_\cap$	0.250	0.791	0.357	0.624
gloss terms	0.741	0.785	0.510	0.613
gloss terms$_\cap$	0.226	0.793	0.275	0.632
lexfnames	0.583	0.792	0.505	0.629
lexfnames$_\wedge$	0.480	0.790	0.440	0.629
linktypes	0.328	0.793	0.365	0.631
linktypes$_\wedge$	0.277	0.792	0.346	0.626
pos	0.146	0.793	0.239	0.633
pos$_\wedge$	0.146	0.793	0.235	0.632
part-of terms	0.372	0.793	0.368	0.635
lexicalized	0.132	0.793	0.213	0.637
part of other	0.132	0.793	0.216	0.636
gloss of other	0.133	0.793	0.214	0.635
hypernym of other	0.132	0.793	0.227	0.627
Roget's Thesaurus-based				
div info	0.679	0.789	0.471	0.629
div info$_\cap$	0.173	0.793	0.283	0.633
Surface level				
affixes	0.200	0.793	0.274	0.637
affixes$_\wedge$	0.201	0.792	0.272	0.635
last letters	0.481	0.792	0.396	0.634
prepositions	0.136	0.793	0.222	0.635
Web 1T-based				
trigrams	0.571	0.790	0.437	0.615
4-grams	0.558	0.797	0.442	0.604

Table 4: Impact of features; cross validation accuracy for *only one feature type* and *all but one feature type* experiments, denoted by 1 and M-1 respectively. $\cap$–features shared by both n_1 and n_2; $\wedge$–n_1 and n_2 features conjoined by logical AND (e.g., n_1 is a 'substance' $\wedge$ n_2 is a 'artifact')

6 Evaluation

6.1 Evaluation Data

To assess the quality of our taxonomy and classification method, we performed an inter-annotator agreement study using 150 noun compounds extracted from a random subset of articles taken from New York Times articles dating back to 1987 (Sandhaus, 2008). The terms were selected based upon their frequency (i.e., a compound occurring twice as often as another is twice as likely to be selected) to label for testing purposes. Using a heuristic similar to that used by Lauer (1995), we only extracted binary noun compounds not part of a larger sequence. Before reaching the 150 mark, we discarded 94 of the drawn examples because they were included in the training set. Thus, our training set covers roughly 38.5% of the binary noun compound *instances* in recent New York Times articles.

6.2 Annotators

Due to the relatively high speed and low cost of Amazon's Mechanical Turk service, we chose to use Mechanical Turkers as our annotators.

Using Mechanical Turk to obtain inter-annotator agreement figures has several drawbacks. The first and most significant drawback is that it is impossible to force each Turker to label every data point without putting all the terms onto a single web page, which is highly impractical for a large taxonomy. Some Turkers may label every compound, but most do not. Second, while we requested that Turkers only work on our task if English was their first language, we had no method of enforcing this. Third, Turker annotation quality varies considerably.

6.3 Combining Annotators

To overcome the shortfalls of using Turkers for an inter-annotator agreement study, we chose to request ten annotations per noun compound and then combine the annotations into a single set of selections using a weighted voting scheme. To combine the results, we calculated a "quality" score for each Turker based upon how often he/she agreed with the others. This score was computed as the average percentage of other Turkers who agreed with his/her annotations. The score for each label for a particular compound was then computed as the sum of the Turker quality scores of the Turkers who annotated the compound. Finally, the label with the highest rating was selected.

6.4 Inter-annotator Agreement Results

The raw agreement scores along with Cohen's κ (Cohen, 1960), a measure of inter-annotator agreement that discounts random chance, were calculated against the authors' labeling of the data for each Turker, the weighted-voting annotation set, and the automatic classification output. These statistics are reported in Table 5 along with the individual Turker "quality" scores. The 54 Turkers who made fewer than 3 annotations were excluded from the calculations under the assumption that they were not dedicated to the task, leaving a total of 49 Turkers. Due to space limitations, only results for Turkers who annotated 15 or more instances are included in Table 5.

We recomputed the κ statistics after conflating the category groups in two different ways. The first variation involved conflating all the TOPIC categories into a single topic category, resulting in a total of 37 categories (denoted by κ* in Table 5). For the second variation, in addition to conflating the TOPIC categories, we conflated the ATTRIBUTE categories into a single category and the PURPOSE/ACTIVITY categories into a single category, for a total of 27 categories (denoted by κ** in Table 5).

6.5 Results Discussion

The .57-.67 κ figures achieved by the Voted annotations compare well with previously reported inter-annotator agreement figures for noun compounds using fine-grained taxonomies. Kim and Baldwin (2005) report an agreement of 52.31% (not κ) for their dataset using Barker and Szpakowicz's (1998) 20 semantic relations. Girju et al. (2005) report .58 κ using a set of 35 semantic relations, only 21 of which were used, and a .80 κ score using Lauer's 8 prepositional paraphrases. Girju (2007) reports .61 κ agreement using a similar set of 22 semantic relations for noun compound annotation in which the annotators are shown translations of the compound in foreign languages. Ó Séaghdha (2007) reports a .68 κ for a relatively small set of relations (BE, HAVE, IN, INST, ACTOR, ABOUT) after removing compounds with non-specific associations or high lexicalization. The correlation between our automatic "quality" scores for the Turkers who performed at least three annotations and their simple agreement with our annotations was very strong at 0.88.

Id	N	Weight	Agree	κ	κ*	κ**
1	23	0.45	0.70	0.67	0.67	0.74
2	34	0.46	0.68	0.65	0.65	0.72
3	35	0.34	0.63	0.60	0.61	0.61
4	24	0.46	0.63	0.59	0.68	0.76
5	16	0.58	0.63	0.59	0.59	0.54
Voted	150	NA	0.59	0.57	0.61	0.67
6	52	0.45	0.58	0.54	0.60	0.60
7	38	0.35	0.55	0.52	0.54	0.56
8	149	0.36	0.52	0.49	0.53	0.58
Auto	150	NA	0.51	0.47	0.47	0.45
9	88	0.38	0.48	0.45	0.49	0.59
10	36	0.42	0.47	0.43	0.48	0.52
11	104	0.29	0.46	0.43	0.48	0.52
12	38	0.33	0.45	0.40	0.46	0.47
13	66	0.31	0.42	0.39	0.39	0.49
14	15	0.27	0.40	0.34	0.31	0.29
15	62	0.23	0.34	0.29	0.35	0.38
16	150	0.23	0.30	0.26	0.26	0.30
17	19	0.24	0.26	0.21	0.17	0.14
18	144	0.21	0.25	0.20	0.22	0.22
19	29	0.18	0.21	0.14	0.17	0.31
20	22	0.18	0.18	0.12	0.10	0.16
21	51	0.19	0.18	0.13	0.20	0.26
22	41	0.02	0.02	0.00	0.00	0.01

Table 5: Annotation results. Id – annotator id; N – number of annotations; Weight – voting weight; Agree – raw agreement versus the author's annotations; κ – Cohen's κ agreement; κ* and κ** – Cohen's κ results after conflating certain categories. Voted – combined annotation set using weighted voting; Auto – automatic classification output.

The .51 automatic classification figure is respectable given the larger number of categories in the taxonomy. It is also important to remember that the training set covers a large portion of the two-word noun compound instances in recent New York Times articles, so substantially higher accuracy can be expected on many texts. Interestingly, conflating categories only improved the κ statistics for the Turkers, not the automatic classifier.

7 Conclusion

In this paper, we present a novel, fine-grained taxonomy of 43 noun-noun semantic relations, the largest annotated noun compound dataset yet created, and a supervised classification method for automatic noun compound interpretation.

We describe our taxonomy and provide mappings to taxonomies used by others. Our inter-annotator agreement study, which utilized non-experts, shows good inter-annotator agreement given the difficulty of the task, indicating that our category definitions are relatively straightforward. Our taxonomy provides wide coverage, with only 2.32% of our dataset marked as other/lexicalized and 2.67% of our 150 inter-annotator agreement data marked as such by the combined Turker (Voted) annotation set.

We demonstrated the effectiveness of a straightforward, supervised classification approach to noun compound interpretation that uses a large variety of boolean features. We also examined the importance of the different features, noting a novel and very useful set of features—the words comprising the definitions of the individual words.

8 Future Work

In the future, we plan to focus on the interpretation of noun compounds with 3 or more nouns, a problem that includes bracketing noun compounds into their dependency structures in addition to noun-noun semantic relation interpretation. Furthermore, we would like to build a system that can handle longer noun phrases, including prepositions and possessives.

We would like to experiment with including features from various other lexical resources to determine their usefulness for this problem.

Eventually, we would like to expand our data set and relations to cover proper nouns as well. We are hopeful that our current dataset and relation definitions, which will be made available via http://www.isi.edu will be helpful to other researchers doing work regarding text semantics.

Acknowledgements

Stephen Tratz is supported by a National Defense Science and Engineering Graduate Fellowship.

References

Ahn, K., J. Bos, J. R. Curran, D. Kor, M. Nissim, and B. Webber. 2005. Question Answering with QED at TREC-2005. In *Proc. of TREC-2005*.

Baldwin, T. & T. Tanaka 2004. Translation by machine of compound nominals: Getting it right. In *Proc. of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*.

Barker, K. and S. Szpakowicz. 1998. Semi-Automatic Recognition of Noun Modifier Relationships. In *Proc. of the 17th International Conference on Computational Linguistics*.

Berger, A., S. A. Della Pietra, and V. J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics* 22:39-71.

Brants, T. and A. Franz. 2006. Web 1T 5-gram Corpus Version 1.1. Linguistic Data Consortium.

Butnariu, C. and T. Veale. 2008. A concept-centered approach to noun-compound interpretation. In *Proc. of 22nd International Conference on Computational Linguistics (COLING 2008)*.

Butnariu, C., S.N. Kim, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale. 2009. SemEval Task 9: The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions. In *Proc. of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20:1.

Crammer, K. and Y. Singer. On the Algorithmic Implementation of Multi-class SVMs In *Journal of Machine Learning Research*.

Downing, P. 1977. On the Creation and Use of English Compound Nouns. Language. 53:4.

Fellbaum, C., editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Finin, T. 1980. The Semantic Interpretation of Compound Nominals. *Ph.D dissertation* University of Illinois, Urbana, Illinois.

Girju, R., D. Moldovan, M. Tatu and D. Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19.

Girju, R. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*.

Girju, R. 2009. The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: a Cross-linguistic Study. In *Computational Linguistics 35(2) - Special Issue on Prepositions in Application*.

Jespersen, O. 1949. A Modern English Grammar on Historical Principles. Ejnar Munksgaard. Copenhagen.

Kim, S.N. and T. Baldwin. 2007. Interpreting Noun Compounds using Bootstrapping and Sense Collocation. In *Proc. of the 10th Conf. of the Pacific Association for Computational Linguistics*.

Kim, S.N. and T. Baldwin. 2005. Automatic Interpretation of Compound Nouns using WordNet::Similarity. In *Proc. of 2nd International Joint Conf. on Natural Language Processing*.

Lauer, M. 1995. Corpus statistics meet the compound noun. In *Proc. of the 33rd Meeting of the Association for Computational Linguistics*.

Lees, R.B. 1960. The Grammar of English Nominalizations. Indiana University. Bloomington, IN.

Levi, J.N. 1978. The Syntax and Semantics of Complex Nominals. Academic Press. New York.

McCallum, A. K. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002.

Moldovan, D., A. Badulescu, M. Tatu, D. Antohe, and R. Girju. 2004. Models for the semantic classification of noun phrases. In *Proc. of Computational Lexical Semantics Workshop at HLT-NAACL 2004.*

Nakov, P. and M. Hearst. 2005. Search Engine Statistics Beyond the n-gram: Application to Noun Compound Bracketing. In *Proc. the Ninth Conference on Computational Natural Language Learning*.

Nakov, P. 2008. Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In *Proc. the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'08).*

Nastase V. and S. Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proc. the 5th International Workshop on Computational Semantics*.

Nastase, V., J. S. Shirabad, M. Sokolova, and S. Szpakowicz 2006. Learning noun-modifier semantic relations with corpus-based and Wordnet-based features. In *Proc. of the 21st National Conference on Artificial Intelligence (AAAI-06).*

Ó Séaghdha, D. and A. Copestake. 2009. Using lexical and relational similarity to classify semantic relations. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009).*

Ó Séaghdha, D. 2007. Annotating and Learning Compound Noun Semantics. In *Proc. of the ACL 2007 Student Research Workshop*.

Rosario, B. and M. Hearst. 2001. Classifying the Semantic Relations in Noun Compounds via Domain-Specific Lexical Hierarchy. In *Proc. of 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01).*

Sandhaus, E. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Spärck Jones, K. 1983. Compound Noun Interpretation Problems. *Computer Speech Processing*, eds. F. Fallside and W A. Woods, Prentice-Hall, NJ.

Turney, P. D. 2006. Similarity of semantic relations. *Computation Linguistics*, 32(3):379-416

Vanderwende, L. 1994. Algorithm for Automatic Interpretation of Noun Sequences. In *Proc. of COLING-94*.

Warren, B. 1978. Semantic Patterns of Noun-Noun Compounds. Acta Universitatis Gothobugensis.

Ye, P. and T. Baldwin. 2007. MELB-YB: Preposition Sense Disambiguation Using Rich Semantic Features. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007).*